J Labour Market Res
DOI 10.1007/s12651-017-0219-3

ARTICLE

Imputation rules for the implementation of the pre-unification education variable in the *BASiD* Data Set

Nicole Gürtzgen[1,2] · **André Nolte**[3]

Accepted: 11 January 2017

Abstract Using combined data from the German Pension Insurance and the Federal Employment Agency (*BASiD*), this study proposes different procedures for imputing the pre-unification education variable in the *BASiD* data. To do so, we exploit information on education-related periods that are creditable for the Pension Insurance. Combining these periods with information on the educational system in the former GDR, we propose three different imputation procedures, which we validate using external GDR census data for selected age groups. A common result from all procedures is that they tend to underpredict (overpredict) the share of high-skilled (low-skilled) for the oldest age groups. Comparing our imputed education variable with information on educational attainment from the Integrated Employment Biographies (*IEB*) reveals that the best match is obtained for the vocational training degree. Although regressions show that misclassification with respect to *IEB* information is clearly related to observables, we do not find any systematic pattern across skill groups.

Keywords Imputation rules · Administrative data · East Germany · Education · Institutions

✉ Nicole Gürtzgen
nicole.guertzgen@iab.de

André Nolte
nolte@zew.de

[1] Institute for Employment Research, Regensburger Str. 104, 90478 Nuremberg, Germany

[2] University of Regensburg, Regensburg, Germany

[3] Centre for European Economic Research, Department of Labour Markets, Human Resources and Social Policy, L 7.1, 68161 Mannheim, Germany

JEL Classification I2 · C81

Imputationsregeln für die Generierung der Bildungsvariable in den *BASiD*-Daten vor der Wiedervereinigung

Zusammenfassung Der vorliegende Beitrag nutzt administrative Daten der Deutschen Rentenversicherung und der Bundesagentur für Arbeit (*BASiD*) um die Bildungsvariable vor der Wiedervereinigung in Ostdeutschland zu imputieren. Hierfür werden rentenrechtliche Bildungsperioden genutzt. Anhand der Bildungsperioden und Informationen zum ostdeutschen Bildungssystem werden drei Imputationsalgorithmen generiert und mit externen DDR Zensusdaten validiert. Alle drei Algorithmen unterschätzten (überschätzten) den Anteil der Hochqualifizierten (Niedrigqualifizierten) unter den älteren Personen. Ein weiterer Vergleich mit Bildungsinformationen der Integrierten Erwerbsbiographie (IEB) zeigt, dass die größte Übereinstimmung mit Personen aus der Gruppe mit einer Berufsausbildung besteht. Regressionsergebnisse innerhalb der IEB Informationen ergeben des Weiteren, dass eine Missklassifizierung mit beobachteten Variablen korreliert ist. Es bestehen jedoch keine systematischen Zusammenhänge zwischen den Bildungsgruppen.

1 Introduction

The use of administrative data sets in economics and especially labour economics has become more and more important for policy evaluation and empirical research (see, e. g., Card et al. 2010 as an example). Apart from their large sample size, such data sets have the advantage of covering

Published online: 05 April 2017

long time periods and offering precise longitudinal information on key variables such as earnings and labour market states. A further strength of most administrative data sets is that they mitigate problems of panel attrition that typically arise with survey data. In this study, we focus on the 0.25% scientific-use-file of the *BASiD* (2007) data set (henceforth referred to as *BASiD*), which supplements a sample of the German Pension Register (VSKT) with information from the Federal Employment Agency.[1] The data provide longitudinal information on individuals' pension-relevant biographies up to the year 2007. Compared to other administrative data such as the *Integrated Employment Biographies* (*IEB*) from the German Federal Employment Agency, *BASiD* encompasses entire individual employment biographies (in general starting with the age of 14).

An important feature of the *BASiD* data set is that it is particularly attractive for studying the Eastern German labour market. While recent studies based on administrative data have been highly constrained to the period after unification (see e. g. Kohn and Antonczyk 2013), much of the economic literature dealing with the labour market in the German Democratic Republic (GDR) has relied on the German Socioeconomic Panel (*GSOEP*) (see e. g. Bird et al. 1994).[2] These survey data provide retrospective GDR information for the years 1988 and 1989. A great advantage of *BASiD* over the *GSOEP* is that it contains full employment biographies of former GDR citizens prior to German unification. However, a shortcoming of *BASiD* is that it fails to provide information on individual covariates before 1992 (exceptions are gender and age), including the entire time period prior to unification. Given that especially information on educational attainment is of major relevance to many labour market applications, this study proposes different procedures to impute the pre-unification education variable in the *BASiD* data.

While the precision of key variables such as earnings is generally considered to be high, educational information from administrative records is often subject to measurement error.[3] In proposing an imputation procedure for the *BASiD* data, our paper is therefore related to the literature dealing with measurement error and missing values (Little and Rubin 2014, Schafer 2010, Manzari 2004). While much of the literature on German administrative data has proposed procedures to eliminate inconsistencies (Huber and Schmucker 2009) and to impute missing values of earnings (Büttner and Rässler 2008) or education (Fitzenberger et al. 2006, Wichert and Wilke 2012), our approach has to deal with a complete lack of educational information prior to unification. Instead of eliminating inconsistencies of an existing variable, we therefore have to provide a rule that indirectly infers educational information from other variables in the data set. We do so by exploiting information on education-related periods that are creditable for the pension insurance (see Table 13 in the Appendix). Combining these periods with information on the educational system in the former GDR, we propose three different imputation procedures. The first one, *IMP1*, attempts to match the six education categories provided by the *IEB* imposing detailed age constraints from the institutional information on the GDR educational system. The second one, *IMP2*, gives up the age constraints and aims to match somewhat broader education categories, into which the six categories in the *IEB* have been typically summarised in many empirical applications. Finally, the third one, *IMP3*, defines the education level based on potential years of education.

We validate our procedures by using external GDR census data provided by the German Statistical Office (see also Steiner 1986 and Maaz 2002). These data allow us to compare the fraction of individuals in specific education-age group cells resulting from our imputations with the corresponding fractions from the census data for comparable cohorts in 1984. The general picture that emerges is that *IMP2* and *IMP3* tend to overpredict the share of medium-skilled workers and underpredict the share of high-skilled for all age groups, whereas they underpredict (overpredict) the share of low-skilled for the younger (older) age groups. The latter is also true for *IMP1*. Compared with *IMP2* and *IMP3*, *IMP1* gives rise to smaller deviations for the share of medium-skilled, whereas it overpredicts the share of high-skilled especially for the younger age groups. Overall, when balancing out the trade off between goodness of fit and data coverage, the more narrowly defined procedure *IMP1* performs best. However, once one re-classifies those with a GDR *Fachschule* degree as medium skilled workers, *IMP2* is to be preferred over *IMP1* and *IMP3*. After discussing potential sources of misclassification, we

[1] *BASiD* is the abbreviation for "Biographiedaten ausgewählter Sozialversicherungsträger in Deutschland". A larger 1%-sample of the pension part of the data (the sample of the German Pension Register – VSKT) is available via on-site access at the Research Data Centre of the German Pension Insurance. An alternative version of the full *BASiD* 1%-sample (including also the data from the Federal Employment Agency) is also available at the Research Data Centre of the Institute for Employment Research (IAB). While the Pension Insurance version provides monthly spell information, the IAB version provides daily information for all pension relevant episodes (see Hochfellner et al. 2012).

[2] A further data set is the German Life History Study that sampled five birth cohorts in the former GDR born between 1929 and 1971 (for a documentation see Goedicke et al. 2004).

[3] A peculiarity of earnings in *BASiD* is the large extent of censoring due to the constant social security contribution limit of 600 Mark, which permits researchers only to observe the bottom of the earnings distribution.

proceed by comparing our education variable prior to unification with educational information from the *IEB* right after unification. To do so, we first improve the *IEB* education information according to the algorithm proposed by Fitzenberger et al. (2006). For comparison purposes, we then follow Wichert and Wilke (2012) by performing a regression analysis in order to identify the importance of observables for a deviation of our imputed educational information from that in the *IEB*. While misclassification with respect to *IEB* information is clearly related to observables, we do not find any systematic pattern across skill groups.

The remainder of the paper is structured as follows. Sect. 2 describes the *BASiD* data and provides descriptive statistics. Sect. 3 describes the educational system of the GDR and discusses the three imputation rules. Sect. 4 validates the imputation results using external census data. Sect. 5 compares the results from our imputations with educational information from the *IEB* right after unification. The final Sect. 6 concludes.

2 The *BASiD* data and descriptive statistics

BASiD Data. The *BASiD* data supplement information from the *German Pension Register* with various data sources from the German Federal Employment Agency. In this contribution we use the scientific use file (*BASiD*-SUF) provided by the German Pension Insurance, which is a stratified random 0.25% sample. This data set samples individuals within the age range of 30 and 67, who have an active pension account in 2007. The data therefore comprise the birth cohorts from 1940 to 1977,[4] leading to an overall sample of about 60,000 individuals. The sample has been drawn in a disproportionate manner and can be made representative using a weighting factor that is part of the data set (for a detailed description see Bönke 2009, Himmelreicher and Stegemann 2008). To identify former GDR citizens, we first select all individuals who prior to monetary union (i. e., the 30th of June 1990) had exhibited at least one monthly spell of pension-relevant activities in the GDR, i. e. those individuals who had gained East German credit points prior to unification. This gives rise to a sample of 11,331 individuals with a total of 2,790,600 monthly spells.

The data provide longitudinal information on individuals' entire pension-relevant biographies up to the year 2007. Individual work histories cover the period from the year individuals were aged 14 until the age of 67. Panel attrition may arise only due to individuals' death or migrating abroad. In Germany, statutory pension insurance is mandatory for all employees in the private and public sector, thus only excluding civil servants and self-employed individuals.[5] In addition, contributions to the pension insurance are paid by the unemployment or health insurance during periods of unemployment and prolonged illness.

As stressed at the outset, the *BASiD* data provide an ideal basis for analysing labour market related questions of former GDR citizens for several reasons: First, it is the only German administrative data source that encompasses full employment biographies. In particular, the *Pension Register* contains information on all periods for which contributions were paid (employment, apprenticeship training, long-term illness, unemployment) as well as periods without contributions, but which were still creditable for the pension insurance. The latter refers to activities for which an individual receives pension credits, such as periods of school or university attendance after the age of 16 and periods of child rearing and caring.

Second, the *BASiD* data is the only individual level data set that contains employment biographies of former GDR citizens before German unification. After unification, former GDR citizens became entitled to transfer their pension-relevant activities to the FRG (Federal Republic of Germany) pension insurance system. For this purpose, the FRG Pension Insurance recorded all periods prior to unification which were creditable for the pension insurance (see above) as well as earnings up to the GDR social security cap. Given that also entitlements from the so called "Freiwillige Zusatzversicherung" could be (partly) transferred to the Western German system, the data cover the full GDR population for whom past pension-relevant periods have been recorded. The pension data therefore allow researchers to track former GDR workers' entire pre- and post-unification employment histories up to the year 2007. Apart from the individual information on pension-relevant activities and earnings, the *Pension Register* provides information on age and gender.

Starting from 1975 in Western and from 1992 in Eastern Germany, employment spells subject to social security contributions from the *Pension Register* can be merged with data from the German Federal Employment Agency, the *IEB* and the *Establishment History Panel* (*BHP*). The *Establishment History Panel* contains information on the establishment's workforce composition, establishment size as

[4] The cohort structure of our data implies that the earliest period in which we observe insured individuals is the year 1954, when those born in 1940 were 14 years old. During the subsequent years younger cohorts successively enter the data set, which gives rise to an increasingly mixed age structure. To ensure representativeness within the selected cohorts in terms of the working-age population's age structure, we have constructed weights based upon administrative population data from the German Federal Statistical Office.

[5] An exception is provided by self-employed individuals who may voluntarily contribute to the Pension Insurance.

well as sector affiliation (see Table 14 in the Appendix). Finally, the *IEB* provide further time varying individual information on blue- or white-collar status, occupational status, educational status (see Table 15 in the Appendix) and an establishment identifier. A shortcoming of the *BASiD* data is therefore that apart from age and gender most covariates (except for those that can be retrieved from the pension-relevant activities listed in Table 13 in the Appendix) are available after unification only, starting with the year 1992.

3 Education system in the GDR

Before we turn to our imputation procedures, we provide some institutional background information on the educational system in the former GDR. Due to their common roots the educational systems in the former GDR and Western Germany (FRG) exhibit some similarities. However, in the course of the GDR's history, the systems considerably diverged with a strong effect on educational outcomes. While in Western Germany a first selection generally took (and still takes) place during primary school after four years of schooling, GDR students used to spend their first eight years together. Thus, completion of 8th grade may be considered the first official school degree as compared with 9th grade in Western Germany. The completion of 10th grade [referred to as "Polytechnische Oberschule" (*POS*)] was the second highest formal degree in the former GDR. As will be discussed below, in the 1950s and 1960s only a minority obtained a *POS* degree, whereas in later years the majority of students was expected to attend school for ten years (Fuchs–Schündeln and Masella 2016). After completion of *POS*, attending high school two more years [referred to as "Erweiterte Oberschule" (*EOS*)] gave rise to a degree equivalent to the Western German "Abitur" after the completion of the 12th grade. However, as a result from an increasing influence of the "state-governed labour force allocation", access to high school became highly limited since the late 1960s and was not only determined by students' performance but also by their own as well as their parents' political orientation (Hegelheimer 1973; Huinink and Solga 1994).

Regarding the quantitative relevance, the *POS* degree was of minor importance in early GDR years. In the 1960s, the fraction of students leaving school with 8th grade or less was about 50%. This share decreased gradually, until eventually completion of *POS* became the most common qualification. At the end of the socialist regime about 80% finished 10th grade (Maaz 2002). Restrictive access to high school resulted in a stable low fraction of students completing high school of slightly above 10% (Solga 2002). This development diverged considerably from that in Western Germany, where in 1989 a much higher share had attained lower educational qualification (30%) and a high school degree (25%) (Maaz 2002).

As to vocational training, the length of an apprenticeship period was strongly determined by prior educational attainment from schooling. In general, vocational training after completing 8th grade took about 3 to 3.5 years, whereas an apprenticeship period following *POS* lasted only two years (Hegelheimer 1973). Moreover, there was also the possibility to combine a three year apprenticeship training with the

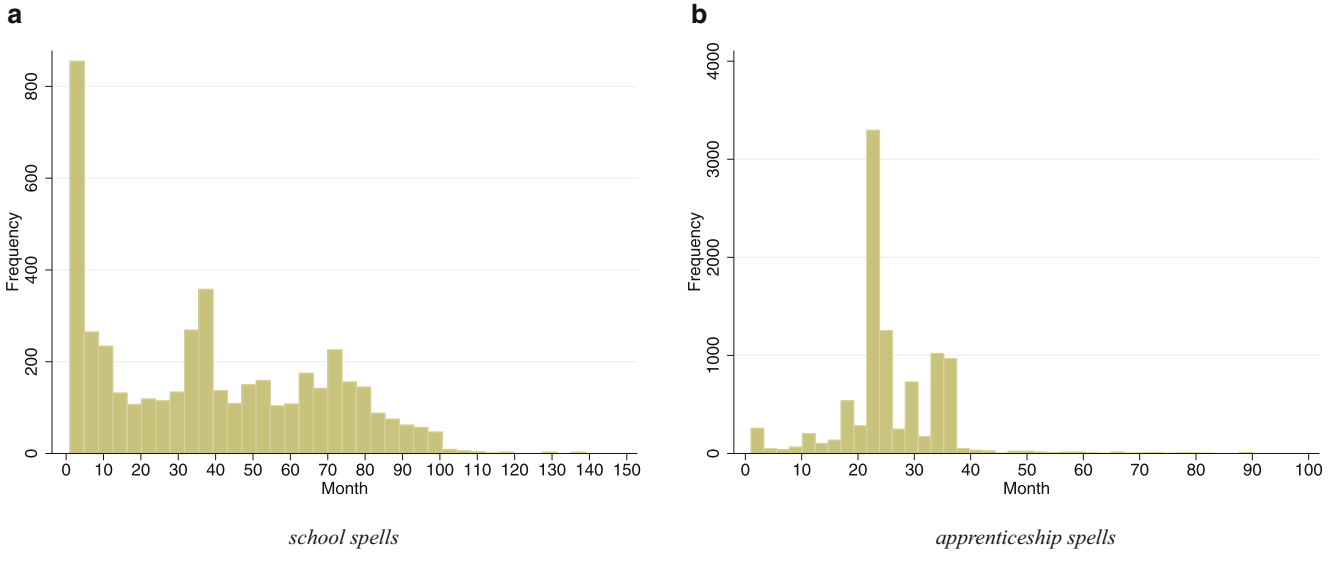

Fig. 1 Distribution of school and apprenticeship spells. (*Source:* BASiD 2007. *Notes:* The figure shows the distribution of the length of schooling (**a**) and apprenticeship spells (**b**). Schooling spells are coded as *ses* = 1. Apprenticeship spells are coded as *ses* = 2. We close interruptions of type 1 and type 3 (see Table 1) if the length of the interruption is less than or equal to six months and if the interruption is not due to a regular employment spell)

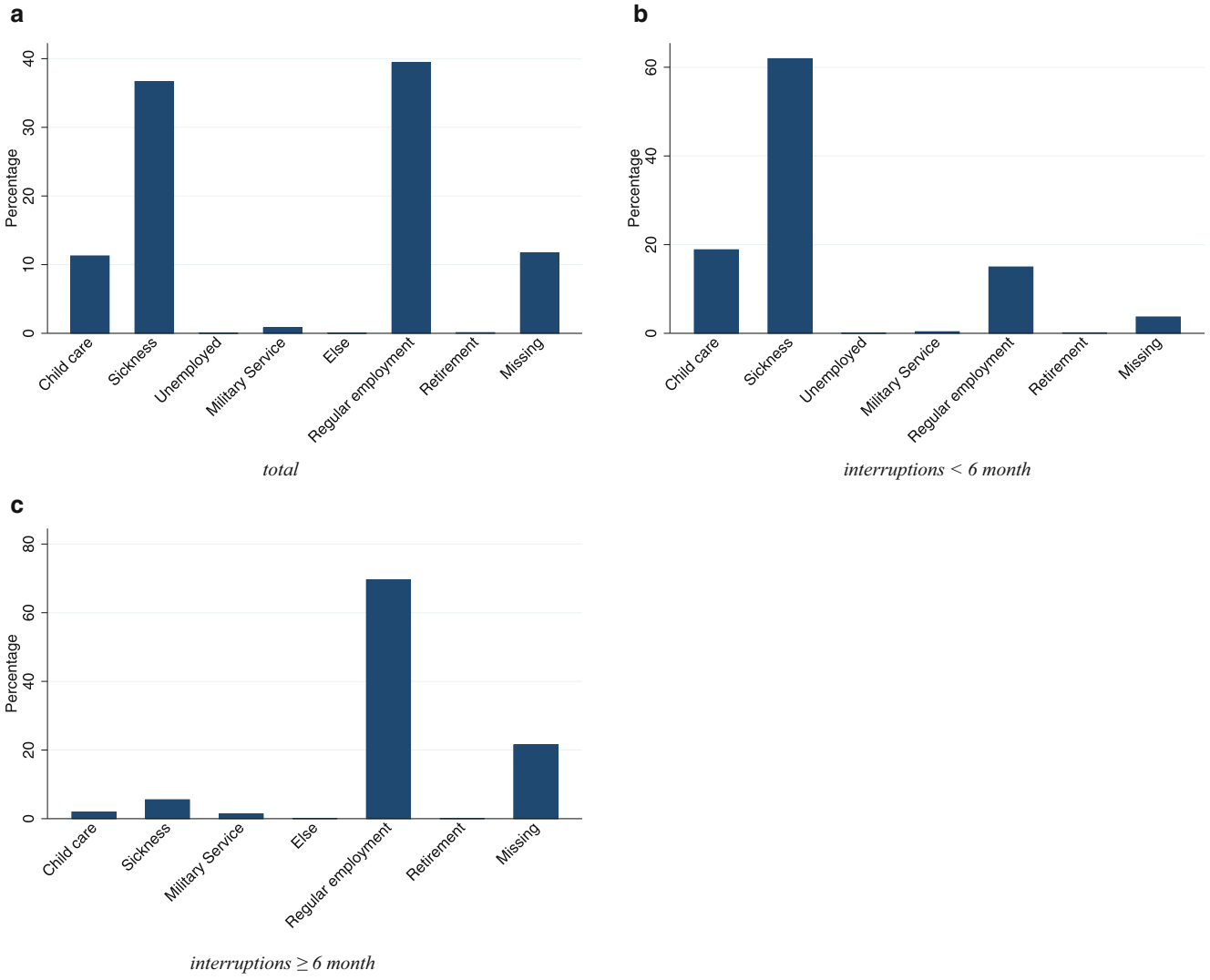

Fig. 2 Socio-economic status during the educational interruption. (*Source:* BASiD 2007. *Notes:* The figure shows the socio-economic status during the educational interruption. Sub-figure **a** pools all available observations. Sub-figure **b** conditions on short interruptions defined as those less than six months. Sub-figure **c** conditions on long interruptions defined as those lasting at least six months or more. See Fig. 3 for a differentiation by type of interruption)

Table 1 Description of the BASiD sample

Panel A: Basic information	
Number of individuals	11,331
... with education interruptions	2,617
... with education interruptions $\geq$ 6 months	1,409
Age at education interruptions	18.3
Panel B: Type of interruption (in%)	
Type 1: school $\rightarrow$ interruption $\rightarrow$ school	29.8
Type 2: school $\rightarrow$ interruption $\rightarrow$ apprenticeship	1.8
Type 3: apprenticeship $\rightarrow$ interruption $\rightarrow$ apprenticeship	40.2
Type 4: apprenticeship $\rightarrow$ interruption $\rightarrow$ school	28.2

Source: BASiD 2007.

completion of a high school degree, which enabled individuals to enter a university afterwards. Given the low share of students with a highest degree of 8th grade or less, in 1987 about 78% of apprenticeship periods exhibited a duration of two years, 11% of 2.5 years and 11% of three years (Fuchs–Schündeln and Masella 2016).

As a further possibility of post-secondary education, individuals could enter a so called *Fachschule*. Admission to a *Fachschule* was either possible after completion of *POS* or, alternatively, after completion of an apprenticeship training. The second type gave rise to a kind of technical university degree (equivalent to that in Western Germany), whereas the first type was closer to a vocational degree (Biermann 2013). The average length of *Fachschule* period took about three years, whereas the length of univer-

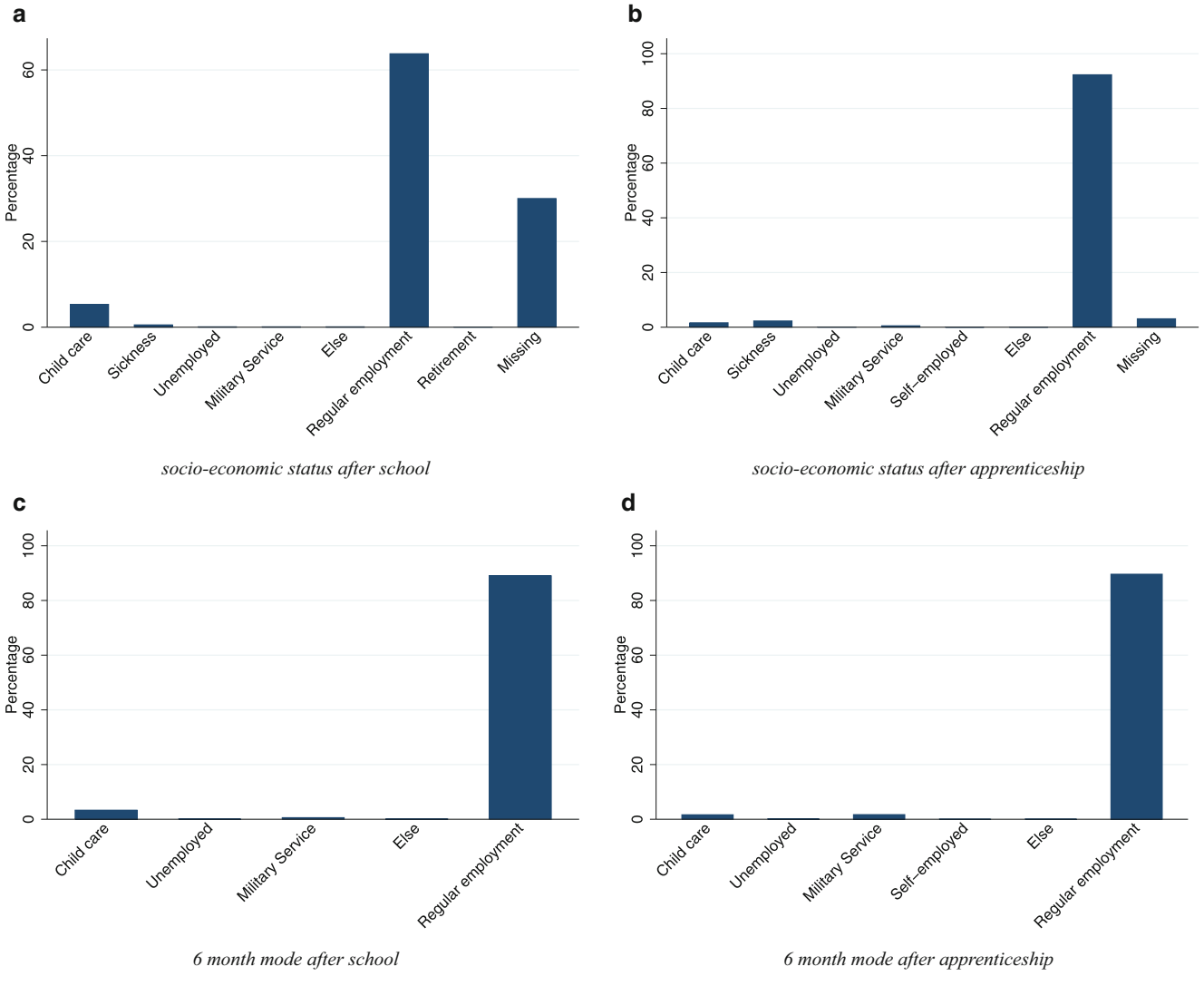

Fig. 3 Socio-economic status after the eduction period. (*Source:* BASiD 2007. *Notes:* The figure shows the socio-economic status after completion of the eduction period. Sub-figures **a** and **b** focus on the first month after the eduction period. Sub-figures **c** and **d** plot the 6 month mode of the socio-economic status)

sity periods was about four years, on average (Krueger and Pischke 1995).

4 Imputation based on educational periods

4.1 Distribution of educational periods

The imputation procedures described in the next section will be basically derived from education-related periods that are creditable for the pension insurance. As spelled out earlier, the *Pension Register* records periods for which contributions were paid (employment, long-term illness, unemployment) as well periods without contributions, but which are still creditable for the pension insurance, e. g. in terms of waiting times. While vocational training periods generally include periods with paid contributions, school and university episodes belong to the second type. By law, the former periods are creditable for the pension insurance of up to 36 months, whereas the latter periods are creditable for the pension insurance for individuals in full time education after the age of 16 for up to 8 years.

To infer information on educational attainment from vocational training and schooling episodes, it is instructive to look at the distributions of these spells. A further relevant issue relates to the question as to what extent individuals experienced educational interruptions and what happened during these interruptions. Fig. 1 displays the distributions of school and apprenticeships spells in our data set.

As can be seen from Fig. 1, a considerable number of individuals exhibits schooling spells of less than six months. There appears to be a further peak around three as well as

six years of schooling. The first peak most likely reflects technical school students, whereas the latter peak relates to university graduates.[6] Apprenticeship spells are, as expected, much more concentrated. Most spells last around two years with a further peak at three years. However, there are many spells that lie in between, indicating either measurement error, errors that may have occurred when transferring pension entitlements to the Western German pension system or uncompleted spells. Breaking down the descriptives by decades (1970s, 1980s, 1990s) shows, however, that the distributions appear to be similar across decades with only some minor differences. In the 1970s, for example, most of the apprenticeship spells lasted three years compared to the 1990s (see Fig. 4 for a detailed graphical illustration over time).

Table 1 shows that among the total number of 11,331 individuals, about 21% (2,414 individuals) experienced some kind of educational interruption. About 46% of the interruptions were rather short. However, 54% of all interruptions lasted longer than six months. Breaking down the descriptions by gender shows that females were slightly more likely (58%) to exhibit educational interruptions. On average, the interruptions started at the age of 18.

Panel B distinguishes four types of interruptions. The figures show that most interruptions occurred between two apprenticeship spells. However, with about 28%, there are also sizeable shares of interruptions between two schooling spells (type 1) and those after a period of vocational training followed by a schooling spell (type 4). The distribution does not substantially differ over time and across gender. For males as well as during the 1980s and 1990s, the third type was slightly more prevalent.

A further interesting information stems from the socio-economic status during the interruption and after the completion of the educational period. This may give some hint of whether individuals interrupted an educational episode due to (regular) employment or due to some other reasons, such as sickness absence.

Fig. 2a shows that most individuals experienced a regular employment spell during the educational interruption (40%) followed by sickness absence with 37.5%. A third important reason is child rearing with about 12%. In 11% of the interruptions, the socio-economic status is missing. The middle part of the figure shows that for short interruptions (defined as those lasting less than six months), sickness absence was the most common reason followed by child rearing (18.8%) and regular employment (17.9%). Individuals experiencing an interruption longer than six months were primarily in regular employment (78%), whereas for about 22% the socio-economic status is not known. Overall, these findings suggest that short-term interruptions are mainly interruptions occurring within one (longer) educational spell.

Panel 1 of Fig. 3 shows the distribution of the socio-economic status in the first month after the educational period. The figure reveals that regular employment is the most frequent status after a school or apprenticeship spell. After schooling periods, the socio-economic status is missing for about 30% of all observations, whereas missing information after an apprenticeship spell is less relevant.

The lower two figures plot the distribution of the main/dominant socio-economic status within the first six months after the educational period. Also at this stage, the main status is regular employment. Interestingly, missing values after schooling episodes do not matter at all as there is no individual with the dominant status being missing after schooling.

4.2 Imputation rules

In what follows, we propose three different imputation procedures, by combining the length of the educational periods described above with information on the educational system in the former GDR. The first procedure, *IMP1*, attempts to match the six education categories provided by the *IEB* imposing detailed age constraints from the institutional information on the GDR educational system. The six categories include (*ND*) no vocational and no high-school degree (henceforth referred to as "No degree"), (*HS*) a high school degree, (*VT*) a completed vocational training, (*VTHS*) a completed vocational training plus high school degree, (*TUD*) a technical university degree and finally, (*UD*) a university degree. The second one, *IMP2*, gives up the age constraints and aims to match the broader education categories, into which the six categories in the *IEB* are typically being summarised in many empirical applications: According to these, (1) low-skilled workers are those without any postsecondary degree, (2) medium-skilled workers have a completed apprenticeship training and (3) high-skilled workers obtained a degree from university or a technical university. Our last procedure, *IMP3*, simply defines these three categories based on potential years of education.[7]

[6] The observation of spells three and six years of schooling may be due to the following reason: Those students who first passed a vocational training degree could enter technical school for three years afterwards. Moreover, students passing EOS (about one year of contribution at the end of EOS) and attending a university for four years afterwards exhibit at least 60 months of contributions.

[7] Before calculating the length of the educational spells, we treat interruptions of such spells in two ways. First, we close interruptions occurring within school and apprenticeship spells of type 1 and type 3 of up to six months if the socio-economic status during the interruptions was not regular employment. This affects 1,194 individuals. Second, we assign the previous school or apprenticeship status for students below 26 if the socio-economic status during the interruption is missing. This affects 1,137 individuals.

Table 2 Imputation Procedure (1)

Category	Characteristics	Criteria	# Individuals
No degree	Age	$<$ 17	778
	First socio-economic spell	employed	
High school (Abitur)	Age	17–18	271
	&		
	School spell	1.0–1.5 years	
Vocational training	Age	$\leq$ 18	3,711
	&		
	Apprenticeship spell	1.5–3.5 years	
Vocational training and high school	Age	19–20	653
	&		
	Apprenticeship spell	2.5–3.5 years	
Technical university degree	Age	19–20	503
	&		
	School spell	2–3.5 years	
	Age	$\geq$ 20	533
	&		
	School spell	1.5–3.5 years	
	Age	$\geq$ 20	375
	&		
	Apprenticeship spell	$>$ 2.5 years	
University degree	Age	$\geq$ 22	688
	&		
	School spell	$\geq$ 3.5 years	

Source: BASiD 2007.
Notes: The total number of individuals under consideration is 11,331. The coverage rate (covered individuals = 7,026/total individuals = 11,331) of Imputation Procedure (1) is 62%. Note that the number of covered individuals is less than the sum over all categories because of transitions to a higher educational level. We allow for interruptions of the school and apprenticeship spells and for continuing the spell's duration after the interruption. See also for a graphical illustration of the procedures and a detailed description of interruptions Appendix B and C. After applying the criteria we extrapolate the educational degree to future spells until we observe a change in individuals' educational status.

Table 3 Imputation Procedure (2)

Category	Characteristics	Criteria	# Individuals
Low-skilled	Age	$\leq$ 17	778
	First socio-economic spell	employed	
Medium-skilled	Apprenticeship	$>$ 1.5 years	7,845
High-skilled	School	$>$ 3 years	1,374

Source: BASiD 2007.
Notes: The total number of individuals under consideration is 11,331. The coverage rate (covered individuals = 9,634/total individuals = 11,331) of Imputation Procedure (2) is 85%. Note that the number of covered individuals is less than the sum over all categories because of transitions to a higher educational level. Cumulative spells allow for interruptions and for continuing the spell's duration after the interruption. After applying the criteria we extrapolate the educational degree to future spells until we observe a change in individuals' educational status.

In general, the three different procedures involve a trade-off between precision and data coverage. *IMP1* bears the potential of losing information on individuals who exhibit school or vocational training spells, but not at the required age. By giving up or loosening the age constraints, *IMP2* and *IMP3* may overcome this loss of observations at the expense of less precision. Table 2 summarises *IMP1* imposing the age constraints (for a graphical summary of the procedures see also Figure C.1. in the Appendix). For this procedure, we need to know at what ages the different degrees were typically completed in the socialist system. As documented by Krueger and Pischke (1995), 8th grade and 10th grade were passed at age 14 and 16, respectively. High school was generally completed at age 18.

According to *IMP1*, individuals are assigned "No degree" if they exhibit a first employment spell and are

younger than 17. Note that in this case the skill status may change once individuals have experienced school or apprenticeship spells according to the rules spelled out below. We choose an age limit of 17 because students passing an apprenticeship of three years after the 8th grade were on average 17 years old. Thus, even in the case of missing values, the first employment spell should be observed at the age of 17 or older. A student with a high school degree (*EOS*) should have completed high school at the age of 18 and should exhibit 12 months of schooling in the data. Thus, individuals having experienced a schooling episode between 1 and 1.5 years at the age of 17 or 18 are assigned a high school degree (category (*HS*) from the *IEB*). To match categories (*VT*) and (*VTHS*) from the *IEB*, we distinguish between those with a completed vocational training after 8th or 10th grade and those combining high school with apprenticeship training. Individuals are assigned to category (*VT*) (8th grade or 10th grade with apprenticeship) if they have experienced a vocational training spell at the age of 18 or younger with the length of the training lasting between 1.5 and 3.5 years.[8]

As already mentioned, a further possibility was to combine a three year apprenticeship training with the completion of a high school degree. Thus, individuals having experienced an apprenticeship spell lasting between 2.5 to 3.5 years at the age of 19 or 20 are assigned an apprenticeship plus high school degree (*VTHS*).

In the *IEB* data, the high-skilled comprise those with a technical university degree and a university degree. In what follows, we will simply match the *IEB* technical university degree with a GDR technical school (*Fachschule*) degree.[9] Our assignment rule relies on the fact that completion of a *Fachschule* took three years, whereas the completion of a university degree took four years on average. To capture the completion of a *Fachschule* degree after *POS*, individuals having experienced a schooling episode of at least 2 up to 3.5 years at the age of 19 to 20 are assigned (*TUD*). The same assignment is made for individuals having experienced a schooling episode of at least 1.5 up to 3.5 years at the age of 20 or older.[10] As we observe a considerable fraction of individuals having experienced an apprenticeship period of at least 2.5 years after the age of 20, those individuals are assigned a *Fachschule* degree as well. This is to account for the possibility that a *Fachschule* degree, which was frequently associated with the completion

Table 4 Qualification structure by age groups – 1984, I

	Age group	Low-skilled	Medium-skilled	High-skilled
IMP1	20–24	8.3	66.3	25.4
	25–29	8.6	58.7	32.7
	30–34	11.9†	62.4†	25.7†
	35–39	20.4	55.6	24.0
	40–44	25.9	58.3†	15.8
IMP2	20–24	2.3	92.3	5.4
	25–29	3.6	81.1	15.4
	30–34	7.9	77.7	14.4
	35–39	13.8†	76.2	10.0
	40–44	23.7	68.8	7.5
IMP3	20–24	5.8	92.3	1.9
	25–29	6.4	86.8	6.8
	30–34	12.2†	80.8	7.0
	35–39	16.3	79.2	4.5
	40–44	32.2	64.1	3.7
Census data 1981	20–24	14.1	74.0	11.9
	25–29	11.5	65.5	23.0
	30–34	11.1	62.5	26.4
	35–39	11.8	60.2	28.1
	40–44	16.5	60.0	23.5
Census data age in 1984	20–24	20.5	74.0	5.5
	25–29	12.5	70.4	17.2
	30–34	11.3	63.6	25.1
	35–39	11.4	61.6	27.0
	40–44	13.2	60.1	26.7

Source: BASiD 2007, weighted statistics.
Notes: Census data cover the residential population in 1981 and are provided as a scientific use file by the German Statistical Office. All statistics are conditional on not being in education at the time of the interview. The first census figures correspond to the year 1981, the interview year. The figures in the lowest panel provide the shares of those who would have been in the respective age groups in 1984. Low-skilled individuals in the census data are individuals without any degree and with a partial completion of a vocational training. Medium-skilled workers include those with a completed vocational training and so-called "Meister", whereas high-skilled workers consist of those with a technical school degree and a university degree. It is assumed that the working population does not substantially differ from the residential population because of zero unemployment. † indicates insignificant differences compared to the census data measured in 1981.

[8] Note that we also assign individuals younger than 17 to category (*VT*) if they have experienced apprenticeship spells of less than three years to account for the possibility of potential underreporting of vocational training spells in the data set. Moreover, we also allow for vocational training spells of more than 36 months of contributions. After consulting the Pension Insurance, they provided us with the information that some so called 'Träger' might still have reported more than 36 months as vocational training, whereas others might have reported employment. Fig. 1 also shows few observations with spells lasting longer than 36 months.

[9] We will discuss the limitations of such an approach below.

[10] Note that for this group we also count schooling spells of less than two years. The reason is that – given the predetermined educational biographies – any experience of a schooling spell at the age of 20 or older is likely to reflect a higher educational degree.

Table 5 Differences among imputation procedures – 1984, part 1

	Differences Q	Individuals N	Coverage C	Q/N	Q/C
Sum of squared difference					
IMP1	687.5	5,698	0.53	0.121	1297.2
IMP2	2242.7	7,556	0.71	0.297	3158.7
IMP3	3551.9	9,027	0.82	0.393	4331.6
Absolute difference					
IMP1	84.1	5,698	0.53	0.015	158.7
IMP2	166.4	7,556	0.71	0.022	234.4
IMP3	204.7	9,027	0.82	0.023	249.6

Source: BASiD 2007, weighted statistics.
Notes: Total number of individuals in 1984: 10,702. The sum of square differences and absolute differences are estimated using the age groups in 1981. The coverage rate of *IMP3* is rather low because for some individuals the requirements were not met at that point in time.

Table 6 Qualification structure by age groups and gender – 1984

	Age	Medium-skilled		Technical School		University	
	group	m	f	m	f	m	f
IMP1	20–25	74.5	60.6	15.6	29.4	0.0	1.2
	25–30	64.7	54.2	17.6	18.6	6.6	10.2
	30–35	66.0	59.5	12.3	11.4	10.3	8.0
	35–40	60.5	51.8	12.0	16.0	10.0	3.5
	40–45	66.9	53.7	9.6	9.8	9.3	2.9
Census data	20–25	81.3	66.3	1.6	18.0	1.3	3.5
1981	25–30	72.1	58.6	6.1	18.8	10.4	11.0
	30–35	67.0	57.8	11.4	20.0	12.8	8.7
	35–40	63.5	56.8	15.4	21.4	12.9	6.3
	40–45	62.3	57.8	14.6	15.4	12.0	4.9

Source: BASiD 2007, weighted statistics.
Notes: Census data cover the residential population in 1981 and are provided as a scientific use file by the German Statistical Office. It is assumed that the working population does not substantially differ from the residential population because of zero unemployment.

of a vocational training, might have been (mis)reported as a vocational training period in the Pension data.[11] Finally, individuals are assigned a university degree (*UD*) if they have experienced a schooling spell of at least 3.5 years at the age of 22 years or older.

Procedure (2) is summarised in Table 3. Low-skilled workers are those assigned "No degree" (see *IMP1*). Medium-skilled workers need to have at least 1.5 years of formal apprenticeship training, whereas high-skilled workers are those with school spells of at least three years.

The third approach relies on potential years of education (*IMP3*). Given that formal unemployment was officially barely present in the GDR[12], the idea of this rule is that the first employment spell should have immediately followed the completion of an educational degree. We define individuals as low-skilled if they are less than 17 years old and are labeled as employed (employment subject to social security contributions excluding apprenticeship periods). Individuals starting employment between 17 and 20 are defined as medium-skilled, whereas high-skilled individuals are those with a first employment spell at the age of 21 to 28. Note that this approach does not account for potential changes in educational attainment over individuals' life courses.[13]

[11] In the Pension Insurance's documentation ("Benutzerhinweise zu den sozialen Erwerbssituationen"), vocational training spells may explicitly cover apprenticeship periods, preparatory vocational training measures as well as technical school ("Fachschule") spells. The latter may also be coded as schooling spells, such that there is some ambiguity with respect to technical school episodes.

[12] See a discussion by Gürtler et al. (1990) on hidden unemployment.

[13] Among the 11,331 individuals in the sample, the fraction exhibiting employment as the first state is 11.3%, out of which about 19% returned to education at some later point in time. This accounts for 2.2% of the whole sample. In total, the procedure generates 1,373 low-skilled, 9,089 medium-skilled and 609 high-skilled individuals. The *IMP3* approach has a coverage rate of 99%.

Table 7 Qualification structure by age groups – 1984, II

	Age group	Low-skilled	Medium-skilled	High-skilled
IMP1	20–24	8.3	90.2	1.5
	25–29	8.6	76.9†	14.5
	30–34	11.9†	74.2	13.9
	35–39	20.4	69.9	9.7†
	40–44	25.9	68.0	6.1
IMP2	20–24	2.3	92.3	5.4
	25–29	3.6	81.1†	15.4
	30–34	7.9	77.7†	14.4
	35–39	13.8†	76.2†	10.0†
	40–44	23.7	68.8	7.5†
IMP3	20–24	5.8	92.3	1.9†
	25–29	6.4	86.8	6.8
	30–34	12.2†	80.8†	7.0
	35–39	16.3	79.2†	4.5
	40–44	32.2	64.1	3.7
Census data	20–24	14.1	83.6	2.3
1981	25–29	11.5	77.8	10.7
	30–34	11.1	78.1	10.8
	35–39	11.8	78.6	9.6
	40–44	16.5	75.1	8.5

Source: BASiD 2007, weighted statistics.
Notes: Census data cover the residential population in 1981 and are provided as a scientific use file by the German Statistical Office. It is assumed that the working population does not substantially differ from the residential population because of zero unemployment. The medium-skilled group now includes former technical school graduates. † indicates insignificant differences compared to the Census data measured in 1981.

5 Qualification structure

This section attempts to externally validate our proposed imputation procedures. To do so, we compare the qualification structures by age groups obtained from our procedures with those from GDR census data from 1981 for the whole residential population. Note that due to the coverage of the pension data and basically zero unemployment, the working population should not substantially differ from the residential population.

The first three panels in Table 4 show the imputed qualification structures distinguished by five age groups, whereas the last two panels display the figures from the census data. Note that our imputed figures refer to the same age groups three years later in 1984 as the cohort structure of our data set allows us to provide figures for the oldest age group (40–44) only from 1984 onwards. For comparison, the first Census panel reports the shares of all individuals who were within the relevant age groups in 1981. The bottom Census panel displays the skill shares of those individuals who would have been in the respective age groups in 1984. The purpose of this comparison is to obtain some information on cohort-specific trends.

The census figures shown in the upper bottom panel indicate that the share of low-skilled workers is U-shaped, whereas the share of medium-skilled workers is decreasing with age. For high-skilled individuals, we observe a kind of inverse U-shaped picture with the largest share of high-skilled individuals in the second oldest age group. For the younger age groups, *IMP1* assigns substantially more individuals to the high-skilled group compared to the census figures and less to the medium and low-skilled category. For the older age groups, it assigns substantially more individuals to the low-skilled and less to the high-skilled category. *IMP2* generates high and low-skilled (medium-skilled) shares that are substantially lower (higher) for most of the age groups than those from the census data. *IMP3* leads to an even more pronounced underprediction of the share of high-skilled in all age groups. This reflects that *IMP3* does not account for completed school episodes at the age of 21 or younger that might have led to a *Fachschule* (technical school degree) after completion of *POS*. Looking at the low-skilled share obtained from *IMP3* suggests that the average age of the first labour market spell increased over time. This is due to the fact that the dominant school degree moved from 8th grade or less (about 30% of school-leavers had less than 8th grade in the 50s and up to 60% had 8th grade) to the 10th grade over that time period (see for a detailed description Solga 2002).

The bottom panel displays the census data of those who would have been in the respective age groups in 1984. This creates substantial deviations for the youngest group, given that some of these individuals were only 17 years old at time of the interview. Because most of the high-skilled individuals were still in the educational system, we observe a higher share of low-skilled and a lower share of high-skilled individuals. This suggests that the comparison is less meaningful for the youngest age group. For the older age groups, however, the pictures does not change substantially. Only those in the age group 40 to 44 in 1984 exhibit lower low-skilled and higher high-skilled shares.

Overall, the discrepancies are non-negligible for each of our proposed procedures. To rank the procedures in terms of the implied deviations from the census data, we calculate the sum of squared differences and the absolute differences between the imputed and the census figures over the different age-skill cells. Table 5 presents the results. *IMP1* results in a sum of squared differences equal to 688, whereas *IMP2* involves a sum of squared differences equal to 2243. Using *IMP3* we obtain a number of 3552.

Based on these numbers, we would prefer the first procedure over the second and the last one. The ranking remains the same by using absolute differences. However, the different procedures capture different numbers of observations.

Table 8 Differences among Imputation Procedures – 1984, part 2

	Differences Q	Individuals N	Coverage C	Q/N	Q/C
Sum of squared difference					
IMP1	492.3	5,698	0.53	0.086	928.9
IMP2	445.5	7,556	0.71	0.059	627.5
IMP3	727.1	9,027	0.82	0.082	886.7
Absolute difference					
IMP1	72.6	5,698	0.53	0.013	137.0
IMP2	67.6	7,556	0.71	0.008	95.2
IMP3	84.7	9,027	0.82	0.009	103.3

Source: BASiD 2007, weighted statistics.
Notes: Total number of individuals in 1989: 11331.

Table 9 Qualification structure by age groups – 1989

	Age group	Low-skilled	Medium-skilled	High-skilled
IMP1	20–24	6.6	90.2	3.2
	25–29	5.4	82.1	12.5
	30–34	8.4	77.8	13.5
	35–39	12.3	75.8	11.9
	40–44	20.7	70.3	9.0
Census data	20–24	14.1	83.6	2.3
1981	25–29	11.5	77.8	10.7
	30–34	11.1	78.1	10.8
	35–39	11.8	78.6	9.6
	40–44	16.5	75.1	8.5

Source: BASiD 2007, weighted statistics.
Notes: Census data cover the residential population in 1981 and are provided as a scientific use file by the German Statistical Office. It is assumed that the working population does not substantially differ from the residential population because of zero unemployment. The medium-skilled group now includes former technical school graduates.

Dividing Q by the number of individuals (or equivalently by the coverage rate), the ranking stays the same. The same conclusion holds when using absolute differences.

What still remains to be resolved is the question as to why we observe these differences, especially for the share of high-skilled. One possible explanation for the strong deviations produced by *IMP2* and *IMP3* could be that the census data are biased towards an over-reporting of high-skilled individuals. An alternative explanation could be that there is a reporting error for school and apprenticeship spells in the administrative data set. Regarding the second explanation (reporting error) we do not think that this is a major problem as especially *IMP2* and *IMP3* should also capture those, whose pension accounts in terms of schooling and apprenticeship episodes might have been incomplete.

Moreover, note that even procedure *IMP1*, which explicitly attempts to distinguish those with a *Fachschule* from those with a university degree, substantially underpredicts (overpredicts) the share of high-skilled especially for the older (younger) age groups. To obtain a more precise picture of potential misclassification sources, we break down the results from *IMP1* by female and male workers as well as by those with a technical school and university degree (see Table 6). The resulting figures show that the way we assigned the education variable performs relatively well for university graduates (for both male and female workers) and medium-skilled male workers as compared to the technical school degree, where large discrepancies can be observed for all age groups.

To account for a potential misclassification of a technical school as a vocational degree, one approach to handle this difficulty could be to re-classify the medium-skilled by assigning all technical school graduates to the medium-skilled. After this re-classification, the high-skilled group would only include university graduates. In terms of the Western German skill categories, such a re-classification could e. g. be justified by the fact that in the GDR individuals could enter a technical school after completion of *POS* (Krueger and Pischke 1995), which would be rather equivalent to a Western German vocational training degree. The results from this re-classification are shown in Table 7. The upper part of Table 7 shows that the share of high-skilled workers decreases and becomes closer to the official data particularly for the older age groups. Thus, treating former *Technical School* graduates as medium-skilled may help to draw a somewhat clearer picture for the high-skilled group. The table also shows again the shares generated by *IMP2* and *IMP3*. After the re-classification, for IMP2 six out of 15 cells are not significantly different from the census data.

After the re-classification, the sum of squared and absolute differences become smaller for all three imputation procedures, with *IMP2* and *IMP3* showing stronger improvements.[14]

The ranking of the procedures changes as well. Based on the Q measure and the number of individuals (coverage rate), *IMP2* is now preferred over *IMP3* and *IMP1*.

Given the substantial deviations for the older age groups that result from all three procedures, we next check whether

[14] Note that for *IMP2* and *IMP3* the improvement only results from re-classifying the census data.

Table 10 Cross tabulation of *IMP1* vs. *IEB*, I

	Degree *IMP1*					
IEB (*IP1*)	Missing	ND	VT	VTHS	TUD	UD
Missing	**33.9**	39.7	31.7	32.5	37.8	25.5
ND	4.7	**14.1**	4.7	3.0	1.5	0.4
VT	56.6	44.7	**61.4**	56.7	43.7	17.2
VTHS	1.6	0.6	1.0	**3.1**	6.9	5.8
TUD	2.0	0.5	1.5	3.0	**15.3**	10.3
UD	1.1	0.1	0.1	1.6	4.5	**40.4**
Observations	4234	778	3566	573	1103	929

Source: BASiD 2007, weighted statistics.
Notes: The category *High school* plays with 1.3% a minor rule and is not presented in the table. Total number of observations is 11,331. The reference point in time for the comparison is January 1993. Abbreviations: ND: no degree, VT: completed vocational training, VTHS: high school with vocational training. TUD: technical university degree, UD: university degree. In the Pension data *TUD* corresponds to a technical school (*Fachschule*) degree.

Table 11 Cross tabulation of *IMP1*, *IMP2* and *IMP3* vs. *IEB*, II

	IMP1			
IEB	Missing	Low-skilled	Medium-skilled	High-skilled
Missing	**33.9**	37.4	31.8	26.8
Low-skilled	4.8	**13.0**	4.1	1.4
Medium-skilled	58.2	47.5	**62.0**	38.0
High-skilled	3.1	2.2	2.1	**33.9**
Observations	4234	926	4139	2032
	IMP2			
IEB	Missing	Low-skilled	Medium-skilled	High-skilled
Missing	**37.9**	39.7	31.4	24.8
Low-skilled	8.8	**14.1**	3.3	0.9
Medium-skilled	46.8	45.4	**62.5**	30.3
High-skilled	6.5	0.8	2.8	**44.1**
Observations	1696	778	7495	1362
	IMP3			
IEB	Missing	Low-skilled	Medium-skilled	High-skilled
Missing	**48.8**	38.1	31.7	25.4
Low-skilled	26.8	**10.9**	3.8	1.4
Medium-skilled	12.2	47.7	**57.8**	33.9
High-skilled	12.2	3.3	6.8	**39.2**
Observations	41	1373	9292	625

Source: BASiD 2007.
Notes: Total number of observations 11,331. The reference point in time for the comparison is January 1993. Abbreviations (IEB): Low-skilled include ND and HS, Medium-skilled include VT and VTHS, High-skilled include TUD and UD.

our procedures are at least able to reproduce the documented decline in the fraction of those leaving school at 8th grade or less. We do this by estimating our skill shares at a later point in time (five years later in 1989). Individuals from the oldest age group (those aged 44) in 1989 were born in 1945 and were potentially available for the labour market in 1960/1961. Given that the share of those leaving school at 8th grade or less was twice as high in the 1950s as compared to the 1960s (Solga 2002), we expect a sharp decline in the predicted low-skilled share especially for older age groups. Table 9 presents the results for *IMP1*.

As can be seen, our expectations are borne out by the figures, which are also in line with the descriptive statistics shown in Solga (2002). (Unreported) results reveal that the share of low-skilled obtained from the other two procedures drop by a similar magnitude. Thus, our procedures at least appear to provide consistent results in terms of the decline of those leaving school at 8th grade or less.

Taken together, our comparison with the census data suggests that all three procedures exhibit non-negligible deviations from our external data source. In order to finalise our decision about which procedure to use, we further compare our imputation results to information provided by the *IEB* subpart of the data set.

6 Comparison with educational information from the *IEB*

We next compare the results from our imputation procedures with educational information from the *IEB*, which can be merged to the *BASiD* data. Even though the *IEB* information starts in 1992, we use January 1993 as a reference point as there is evidence that education information is more reliable from 1993 onwards. With this comparison we have to keep in mind that the *IEB* education information may be subject to measurement error as well. To mitigate this issue, we correct the *IEB* education information using the imputation algorithm *IP1* proposed by Fitzenberger et al. (2006). The authors suggest three different imputation rules without a strict order. The idea behind their rules is based on the assumption that individuals cannot lose their educational degrees. In what follows we use *IP1*, since according to Wichert and Wilke (2012) imputation procedure 1 (*IP1*) leads to a stronger reduction in measurement error.

We perform this exercise for all of our imputation procedures. Table 10 first cross tabulates the results from procedure *IMP1* using the categories described in Table 2 with education information from the *IEB*.

Table 10 shows that the best match is obtained for the vocational training (*VT*) category with a fraction of 61% receiving this category from both our imputation procedure *IMP1* and the *IEB*. In contrast, among those assigned a vocational training plus high school degree (*VTHS*) in the Pension data, over 50% exhibit only a vocational training degree (*VT*) in the *IEB*. Note that this may either reflect that our imputation procedure wrongly assigns a vocational plus high school degree or, alternatively, that the GDR high school degree has either not been reported or recognised by Western German employers.

Moreover, the *IEB* comparison also produces large deviations for those assigned no degree in the Pension data (*ND*), who mostly exhibit a vocational training (*VT*) in the *IEB*. A potential explanation would be that our defined rules from procedure *IMP1* might have changed over time, such that older workers could have completed an apprenticeship within a shorter duration. If this was the case, the deviation of *IMP1* from *IEB* information should be mitigated using *IMP2*, as this procedure requires only a cumulative apprenticeship spell of at least 1.5 years. The second panel of Table 11 shows that this does not account for the observed deviation in Table 10. In Table 11 the share of those assigned no degree *ND* in the Pension data, who exhibit a vocational training (*VT* and *VTHS*) in the *IEB*, remains basically the same for *IMP2*. Note that the share of low-skilled decreased substantially for all age groups between 1984 and 1989, such that over-reporting a low-skilled status is unlikely to explain the deviation for the low-skilled in Table 10. An alternative explanation would be vocational on-the-job training, which generally cannot be ruled out as a source of any deviation between the (imputed) pension data and *IEB* information.[15]

Given the problem with correctly assigning the technical school [or technical university degree (*TUD*)], over 50% of those assigned *TUD* in the Pension data exhibit only a vocational training degree (*VT* or *VTHS*) in the *IEB*. This highlights again the difficulty in distinguishing between a GDR technical school and a vocational degree. Note that this misclassification might also reflect the fact that educational degrees obtained during GDR times might have not been recognised by Western German employers after unification. The overlap of *TUD* with a university degree (*UD*) is only moderate with 4.5%.[16] From the 929 individuals who were assigned *UD* in the Pension data, about 40% exhibit also a *UD* in the *IEB*. 25% have missing values in the *IEB* and about 17% intersect with vocational training (*VT*).

The cross tabulation of the results from *IMP2* and *IMP3* with education information from the *IEB* is shown in Table 11. As these procedures target three education categories – low-skilled, medium-skilled and high-skilled, we also provide the shares for the first imputation procedure *IMP1* using the three categories in the upper panel.

The picture that emerges from Table 11 is that all three procedures show again very similar patterns. The best match is obtained for the medium-skilled, whereas the worst match results for the low-skilled.[17] Overall, for the 11,331 individuals in the *BASiD* data set we obtain the highest number of missing values (4,234) for procedure *IMP1*. The number of missing values in the *IEB* (3,643) variable is lower compared to *IMP1* but twice as large compared to *IMP2*, indicating a substantial gain in information resulting from

[15] We also performed the same analysis with individuals who were unemployed in 1993, with the result of less overlap.

[16] This again suggests that re-classifying a *Fachschule* degree as medium-skilled would potentially reduce measurement error issues and misclassification.

[17] Note that misclassification also occurs when comparing different sources from the *IEB*. Wichert und Wilke (2012) compare job seekers' histories (BewA) with data from the Employment Register (BeH) and obtain high misclassified results (Wichert and Wilke 2012). For example, the match between BewA and BeH data for the technical school degree variable is about 36% after correcting the variable by using the imputation algorithm.

Table 12 Marginal effects of a logit regression

Dependent variable:	Deviation from *IEB* information of ...			
	(1)	(2)	(3)	(4)
	missing	Low-skilled	Medium-skilled	High-skilled
Female	**−0.057** (0.016)	0.002 (0.039)	0.002 (0.014)	**0.172** (0.036)
Age group				
< 25 (ref.)				
25–30	−0.044 (0.047)	0.062 (0.064)	**−0.082** (0.018)	**0.168** (0.075)
30–35	0.008 (0.027)	0.054 (0.059)	−0.008 (0.020)	**0.167** (0.068)
35–40	−0.003 (0.029)	0.049 (0.060)	−0.002 (0.020)	0.136 (0.071)
40–45	−0.026 (0.038)	−0.010 (0.045	−0.013 (0.019)	0.067 (0.073)
45–50	−0.005 (0.030)	−0.000 (0.040)	−0.016 (0.021)	0.091 (0.080)
Firm size				
50–199 (ref.)				
below 20	−0.062 (0.071)	0.079 (0.051)	**0.091** (0.026)	**0.266** (0.063)
20–49	−0.083 (0.089)	−0.138 (0.081)	0.054 (0.029)	−0.016 (0.068)
200–999	−0.025 (0.050)	**−0.116** (0.054)	−0.002 (0.023)	0.078 (0.055)
1000 and above	−0.018 (0.048)	0.053 (0.058)	**0.125** (0.027)	0.090 (0.060)
Economic sector				
Construction (ref.)				
Agrar	0.016 (0.009)	−0.039 (0.059)	−0.025 (0.026)	–
Energy/mining	**−0.463** (0.167)	−0.020 (0.130)	**0.283** (0.076)	**0.258** (0.101)
Manufacturing	–	−0.072 (0.079)	0.001 (0.033)	−0.152 (0.093)
Wholesale	−0.213 (0.159)	0.126 (0.064)	0.086 (0.057)	**0.254** (0.109)
Traffic/communic.	−0.276 (0.199)	0.135 (0.078)	0.055 (0.082)	0.163 (0.139)
Banking/insurance	–	0.080 (0.148)	0.261 (0.156)	−0.041 (0.145)
Other services	−0.141 (0.116)	0.111 (0.189)	0.303 (0.165)	0.001 (0.092)
Non-profit	−0.185 (0.281)	**0.155** (0.068)	0.227 (0.215)	0.067 (0.129)
Public sector	−0.174 (0.175)	0.155 (0.193)	0.274 (0.240)	0.136 (0.093)
Predicted prob.	0.892	0.832	0.224	0.465
Log. likelihood	−135.4	−254.6	−1993.4	−548.7
Observations	555	563	4184	789

Source: BASiD 2007.
Notes: Robust standard errors are in parentheses. Bold numbers represent significance on at least the 5% level.

our imputation procedures.[18] As mentioned earlier, there exists generally a trade-off between precision and coverage in terms of missing values. Given that the main diagonal values are largest for *IMP2*, which simultaneously reduces the number of missing values by almost 50%, procedure *IMP2* seems to provide a quite reasonable compromise between matching *IEB* information and data coverage.

To analyse whether any deviation from *IEB* information is systematically related to observables, we next perform a logit analysis using the results from procedure (*IMP2*) for the year 1993. The dependent variable is one if the educational degree assigned in the Pension data deviates from the educational information in the *IEB* and is zero otherwise. Table 12 presents the estimated marginal effects for the probability of misclassification in the pooled sample separately for each skill group. While the results suggest that misclassification is related to some observables, there appears to be no systematic pattern across skill groups. In particular, the marginal effects of the covariates vary considerably across skill groups. Being female increases the probability of being misclassified by about 17 p. p. for the high-skilled. Individuals in the age group 25 to 35 have a higher probability of being misclassified if they are assigned high-skilled in the Pension data [Column (4)], whereas the coefficient is negative for the medium-skilled category. Individuals employed by smaller establishments are more likely to be misclassified in all skill groups than in the reference group of medium-sized establishments (not significant for low-skilled workers). Individuals employed by large establishments exhibit a higher probability of misclassification only if they are assigned medium-skilled [Column (3)], whereas the marginal effects for low and high-skilled in the Pension data [Column (2) and (4)] are not

[18] Missing values in the IEB part of the data set are with 4,204 missing entries substantially higher in 1992.

significant. The signs of the marginal effects of different industry affiliations do not reveal any systematic pattern either and also vary greatly across skill groups.

7 Conclusions

The *BASiD* data set provides the only available data source that contains full employment biographies of former GDR citizens prior to German unification. However, a shortcoming of *BASiD* is that it fails to provide information on individual covariates prior to unification (exceptions are gender and age). Given that especially information on educational attainment is of major relevance to many labour market applications, this study proposes different procedures to impute the pre-unification education variable in the *BASiD* data.

Our proposed procedures exploit information on education related periods that are creditable for the pension insurance. Combining these periods with information on the educational system in the former GDR, we investigate three different imputation procedures. The first one, *IMP1*, attempts to match the six education categories provided by the *IEB* imposing detailed age constraints from the institutional information on the GDR educational system. The second one, *IMP2*, gives up the age constraints and aims to match somewhat broader education categories, into which the six categories in the *IEB* have been typically summarised in many empirical applications. Finally, the third one, *IMP3*, defines the education level based on potential years of education.

We validate our procedures by using external GDR census data provided by the German Statistical Office. These data allow us to compare the fraction of individuals in specific education-age group cells resulting from our imputations with those from the census data for comparable cohorts in 1984. The general picture that emerges is that *IMP2* and *IMP3* tend to overpredict the share of medium-skilled workers and underpredict the share of high-skilled for all age groups, whereas they underpredict (overpredict) the share of low-skilled for the younger (older) age groups. The latter is also true for *IMP1*. Compared with *IMP2* and *IMP3*, *IMP1* gives rise to smaller deviations for the share of medium-skilled, whereas it overpredicts the share of high-skilled especially for the younger age groups. Overall, when balancing out the trade off between goodness of fit and data coverage, the more narrowly defined procedure *IMP1* performs best.

Finally, a comparison of our (imputed) education information prior to unification with educational information from the *IEB* right after unification suggests that the best fit is obtained for those assigned a vocational training degree in the Pension data. The largest discrepancy is observed for those assigned a technical university degree in the Pension data of whom a large fraction (over 50%) exhibits only a vocational training degree in the *IEB*. This highlights again the difficulty in distinguishing between a GDR *Fachschule* and a vocational training degree. A simple approach to handle this difficulty could be to redefine the medium-skill category by assigning all technical (school) university graduates to the medium-skilled, such that the high-skilled group would only include university graduates. After doing so, *IMP2* would be preferred over *IMP1* and *IMP3*. Given that *IMP2* simultaneously gives rise to the least number of missing values and provides the best match to the *IEB* information, our proposed procedure *IMP2* seems to provide a quite reasonable compromise between matching external information and data coverage.

Acknowledgements We would like to thank Sebastian Butschek, Laura Pohlan and three anonymous referees for helpful comments and suggestions. We further thank Maria Bidenko and Vanessa Lindenmaier for providing excellent research assistance.

Appendix

Appendix A: Data description

Table 13 Description of individual employment history variables gained from the *Pension Register*

Variable[1)]	Definition	SES Coding[2)]
EMPLOYMENT	Employment spells include periods of employment subject to social security contributions and (after 1998) marginal employment.	13
UNEMPLOYMENT	Unemployment spells include periods of unemployment with and without transfer receipt (only FRG).[3)]	6, 7, 8
NON-EMPLOYMENT	Non-employment spells include periods of child raising, care giving as well as periods with missing information on the employment status.	3, 4
ILLNESS	Illness spells include periods of long-term illness (FRG > 6 weeks; GDR > 4 weeks before 1984, no minimum restriction afterwards).	5
TRAINING	Training spells include periods of school or university attendance after the age of 16 and periods of training and apprenticeship.	1, 2

[1)] Note that the recorded pre-unification pension activity histories are less precise than the post-unification histories. The reason is that the transfer of the activities was mainly based on former GDR citizens' social security cards. These cards record the number of months of employment, illness and maternity leave during a particular year, but do not allow for tracking these spells on a monthly basis. As a result, compared to the pension spells after Unification, which provide exact monthly information on all pension relevant activities, information on the incidence of pre-unification employment, illness and maternity leave spells is available only on an annual basis.

[2)] Further possible states of the SES variable are: Military service (SES = 9), Retirement (SES = 15) and "Else" (SES = 12).

[3)] A spell of unemployment in the *Pension Register* requires individuals to be registered as unemployed *and* to obtain public transfers. The latter include benefits such as unemployment insurance, and – prior to 2005 – the means-tested social assistance and unemployment assistance benefits. After 2004, unemployment and social assistance were merged into one unified benefit, also known as 'unemployment benefit II' (ALG II). As the latter targets only employable individuals, a spell involving the receipt of ALG II automatically fulfills the requirements to be recorded as unemployed in the *Pension Register*. Prior to 2005, spells with social assistance benefits fulfill the above requirements only if individuals were registered as unemployed. Otherwise they are recorded as non-employment spells. As a consequence, the *Pension Register* does not permit a consistent definition of un- and non-employment prior to and after 2005.

Table 14 Definition of establishment characteristics gained from the *Employment Statistics Register*

Variable	Definition/categories:
Establishment size	Size < 20
	$20 \leq$ Size < 50
	$50 \leq$ Size < 200
	$200 \leq$ Size < 1000
	Size ≥ 1000
Workforce composition	Share of employees younger than 30 years
	Share of employees older than 50 years
	Share of low-skilled employees
	Share of female employees
Sector affiliation	Agriculture/forestry
	Mining and manufacturing
	Energy/water supplies
	Construction
	Wholesale and retail trade
	Transport and communication
	Financial intermediation
	Other service activities
	Public administration

Table 15 Definition of individuals characteristics

Variable/categories	Definition
GDR-spell	GDR spells are identified based on the
	regional origin (*Beitrittsgebiet*) of the
	pension contributions
Educational status 6 categories	
NO DEGREE (*ND*)	No high school, no vocational degree
HIGH SCHOOL	High school degree (Abitur)
VOC. TRAINING (*VT*)	Completed vocational training
VOC. TRAINING + HIGH SCHOOL (*VTHS*)	Completed vocational training plus
	high school
TECH. UNIVERSITY (*TUD*)	Fachschule or Technical University
	Degree
UNIVERSITY (*UD*)	University degree
Educational status 3 categories	
LOW-SKILLED	No degree or high school degree
MEDIUM-SKILLED	Completed vocational training
HIGH-SKILLED	Technical college degree or university degree
Origin of credit points	
"Beitrittsgebiet" (GDR)	Variable RCEG = 6
	and RTVS = 5 or 6 before unification

Appendix B: Distribution of education spells and interruptions

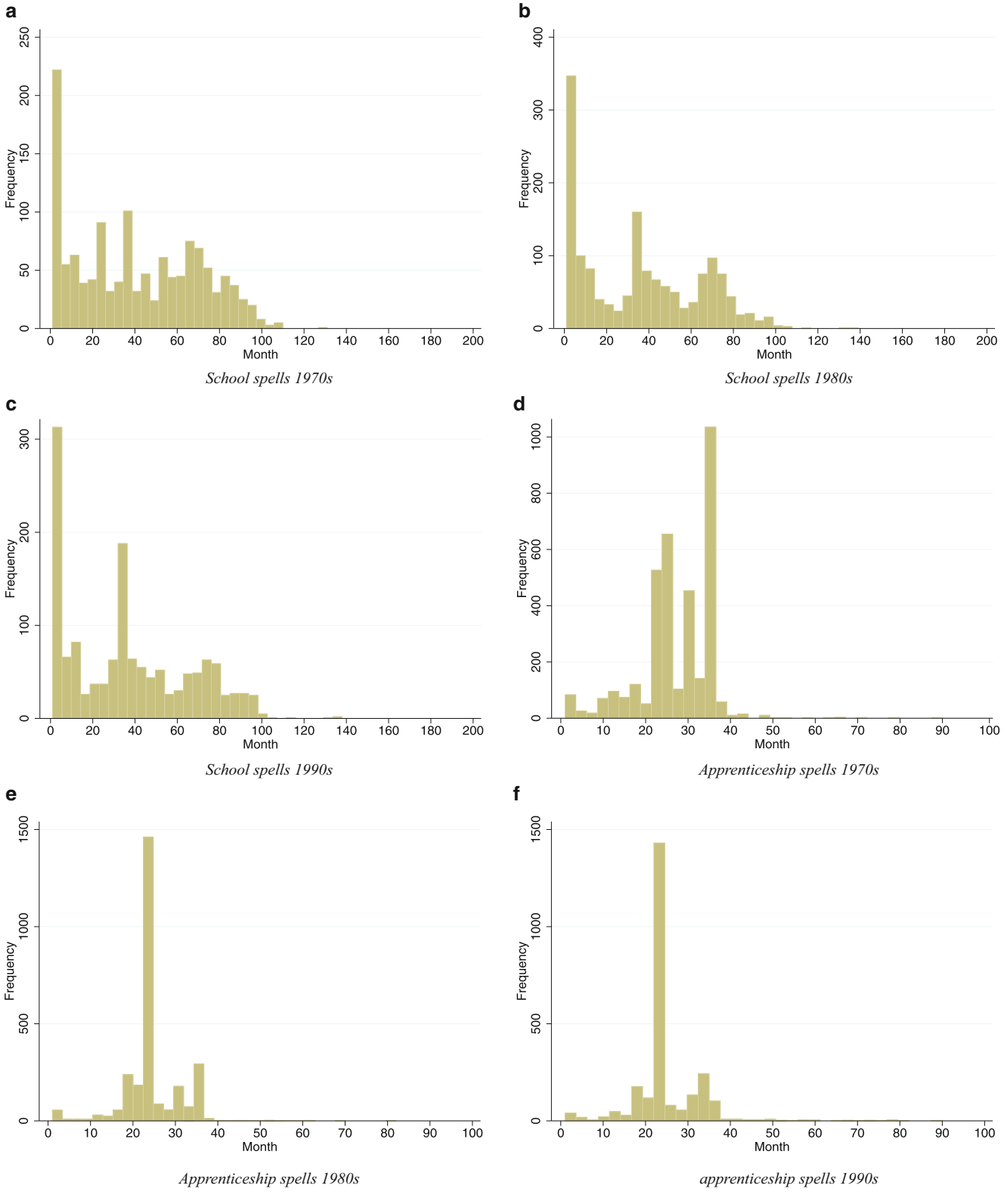

Fig. 4 Distribution of school and apprenticeship spells by period. (*Source:* BASiD 2007. *Notes:* The figure shows the distribution of the length of school (**a–c**) and apprenticeship spells (**e–f**). School spells coded as *SES* = 1. Apprenticeship spells are coded as *SES* = 2. We close interruptions of type 1 and type 3 if the length of the interruption is less than 12 months and if the interruptions do not occur due to regular employment spells)

Appendix C: Graphical illustration of *IMP 1*

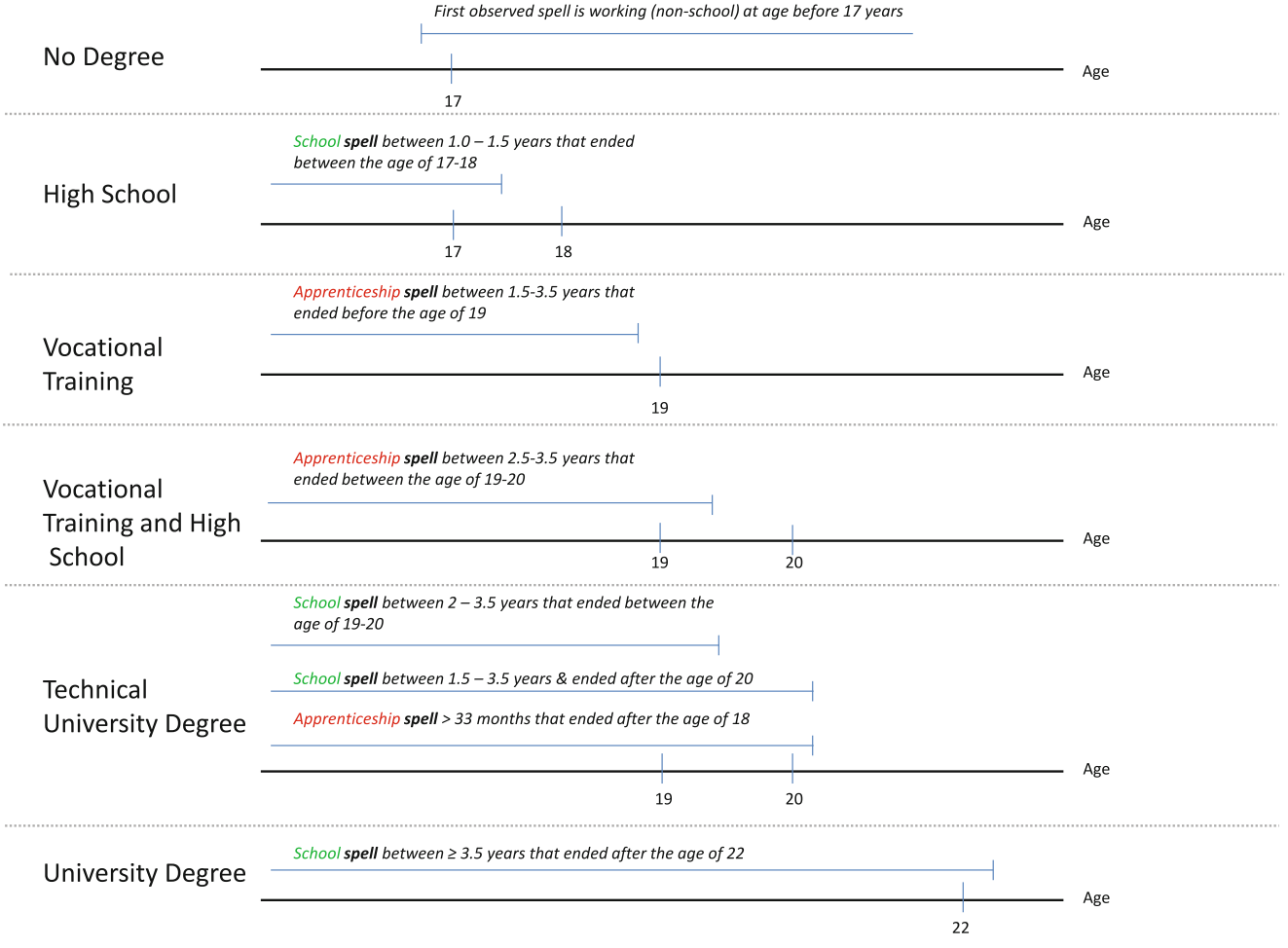

Fig. 5 Graphical Illustration of *IMP1*. (*Notes:* The graph illustrates the imputation rules for *IMP1*. The rules are derived by combining the length of schooling and apprenticeship episodes with information on the educational system. The age thresholds refer to the end of the schooling and apprenticeship episodes)

References

Biermann, H.: Berufsausbildung in der DDR: Zwischen Ausbildung und Auslese. Springer, Berlin (2013)

Bird, E.J., Schwarze, J., Wagner, G.G.: Wage effects of the move toward free markets in East Germany. Ind Labor Relat Rev **47**(3), 390–400 (1994)

Bönke, T.: Gekappte Einkommen in prozessgenerierten Daten der Deutschen Rentenversicherung – Ein paretobasierter Imputationsansatz. DRV-Schriften, vol. 55., pp 214–230 (2009)

Büttner, T., Rässler, S.: Multiple imputation of right-censored wages in the German IAB employment sample considering heteroscedasticity. Technical report. IAB Discuss Pap **44**, 22 (2008)

Card, D., Kluve, J., Weber, A.: Active labour market policy evaluations: A meta-analysis. Econ J **120**(548), F452–F477 (2010)

Fitzenberger, B., Osikominu, A., Völter, R.: Imputation rules to improve the education variable in the IAB employment subsample. J Appl Soc Sci Stud **126**(3), 405–436 (2006)

Fuchs–Schündeln, N., Masella, P.: Long-lasting effects of socialist education. Rev Econ Stat **98**(3), 428–441 (2016)

Goedicke, A., Lichtwardt, B., Mayer, K.U.: Dokumentationshandbuch Ostdeutsche Lebensverläufe im Transformationsprozeß: LV-Ost Nonresponse. Max-Planck-Institut für Bildungsforschung, Berlin (2004)

Gürtler, J., Ruppert, W., Vogler–Ludwig, K.: Verdeckte Arbeitslosigkeit in der DDR. CESifo Group, München (1990). Nr. 100219900050000

Hegelheimer, A.: Berufsausbildung in der DDR – Versuch einer Einschätzung. Gewerksch Monatsh **24**(3), 189–193 (1973)

Himmelreicher, R., Stegemann, M.: New possibilities for socio-economic research through longitudinal data from the Research Data Centre of the German Federal Pension Insurance. J Appl Soc Sci Stud **128**(4), 647–660 (2008)

Hochfellner, D., Müller, D., Wurdack, A.: Biographical data of social insurance agencies in Germany – Improving the content of administrative data. Schmollers Jahrb **132**, 443–451 (2012)

Huber, M., Schmucker, A.: Cleansing procedures for overlaps and inconsistencies in administrative data. The case of length of unemployment in German labour market data. Hist Soc Res **34**, 230–241 (2009)

Huinink, J., Solga, H.: Occupational opportunities in the GDR: A privilege of the older generations? Z Soziol **23**(3), 237–253 (1994)

Kohn, K., Antonczyk, D.: The aftermath of reunification. Econ Transition **21**(1), 73–110 (2013)

Krueger, A.B., Pischke, J.-S.: A comparative analysis of East and West German labor markets: before and after unification. In: Freeman und Katz: differences and changes in wage structures, pp. 405–446. University of Chicago Press, Chicago and London (1995)

Little, R.J.A., Rubin, D.B.: Statistical analysis with missing data. John Wiley & Sons, Hoboken, New Jersey (2014)

Kai Maaz. Ohne Ausbildungsabschluss in der BRD und DDR: Berufszugang und die erste Phase der Erwerbsbiographie von Ungelernten in den 1980er Jahren. *Working Paper of the Independent Research Group of Max-Planck-Institute for Educational Research No. 3/2002*, No. 3/2002, 2002.

Manzari, A.: Combining editing and imputation methods: an experimental application on population census data. J R Stat Soc Ser A **167**(2), 295–307 (2004)

Schafer, J.L.: Analysis of incomplete multivariate data. CRC press, Boca Raton London (2010)

Solga, H.: Jugendliche ohne Schulabschluss und ihre Erwerbsbiografien. In: Schnabel (ed.) Das Bildungswesen in der Bundesrepublik Deutschland. Strukturen und Entwicklungen im Überblick. Rowohlt Taschenbuch Verlag, Reinbek (2002)

Steiner, I.: Struktur der Allgemeinausbildung und Berufsausbildung der Wohnbevölkerung in der DDR – Berufs- und Bildungsweglaufbahnen von Schulabsolventen. Akademie der Pädagogischen Wissenschaft der DDR, Berlin (1986)

Wichert, L., Wilke, R.A.: Which factors safeguard employment?: An analysis with misclassified German register data. J R Stat Soc Ser A **175**(1), 135–151 (2012)

Nicole Gürtzgen has been head of the IAB research department "Labour market processes and institutions" since October 2015. She also holds a professorship in economics with a focus on labour market research at the University of Regensburg. Prior to her appointment at IAB she held a position as a Senior Researcher at the Centre of European Economic Research (ZEW) in Mannheim. From 1990 to 1996, she studied mathematics and economics at the Universities of Duisburg and Heidelberg. After graduating from the University of Heidelberg with a diploma in economics, she completed her doctoral dissertation at the University of Rostock in April 2002. In 2008, she finished her postdoctoral thesis (Habilitation) at the University of Mannheim.

André Nolte is a PhD student of Economics at the Centre for European Economic Research (ZEW) and the University of Mannheim. He joint the department of "Labour Markets, Human Resources and Social Policy" in September 2012. His research examines empirical economics, especially labour economics, public economics and applied microeconometrics. Prior to his doctoral studies he studied economics at the University of Kassel, Rutgers University and the European Business School in Dublin.